Generating an Entailment Corpus from News Headlines

John Burger, Lisa Ferro

The MITRE Corporation
202 Burlington Rd.
Bedford, MA 01730, USA
{john,lferro} @ mitre.org

Abstract

We describe our efforts to generate a large (100,000 instance) corpus of textual entailment pairs from the lead paragraph and headline of news articles. We manually inspected a small set of news stories in order to locate the most productive source of entailments, then built an annotation interface for rapid manual evaluation of further exemplars. With this training data we built an SVM-based document classifier, which we used for corpus refinement purposes—we believe that roughly three-quarters of the resulting corpus are genuine entailment pairs. We also discuss the difficulties inherent in manual entailment judgment, and suggest ways to ameliorate some of these.

1 Introduction

MITRE has a long-standing interest in robust text understanding, and, like many, we believe that adequate progress in such an endeavor requires a well-designed evaluation methodology. We have explored in great depth the use of human reading comprehension exams for this purpose (Hirschman et al., 1999, Wellner et al., 2005) as well as TREC-style question answering (Burger, 2004).

In this context, the recent Pascal RTE evaluation (Recognizing Textual Entailment, Dagan et al., 2005) captured our interest. The goal of RTE is to assess systems' abilities at judging semantic entailment with respect to a pair of sentences, e.g.:

- *Fred spilled wine on the carpet.*
- *The rug was wet.*

In RTE parlance, the antecedent sentence is known as the *text*, while the consequent sentence is known as the *hypothesis*. Simply put, the challenge for an RTE system is to judge whether the text entails the hypothesis. Judgments are Boolean, and the primary evaluation metric is simple accuracy, although there were other, secondary metrics used in the evaluation.

The RTE organizers provided 567 exemplar sentence pairs. This is adequate for system development, but not for the application of large-scale statistical models. In particular, we wished to cast the problem as one of statistical alignment as used in machine translation. MT systems typically use millions of sentence pairs, and so we decided to find or generate a much larger corpus. This paper describes our efforts along these lines, as well as some observations about the problems of annotating entailment data. In Section 2 we describe our initial search for an entailment corpus. Section 3 briefly describes an annotation interface we devised, as well as our efforts to refine our corpus. Section 4 explains many of the issues and problems inherent in manual annotation of entailment data.

2 Finding Entailment Data

In our study of the Pascal RTE development corpus, we found that a considerable majority of the TRUE pairs exhibit a stronger relationship than entailment; namely, the hypothesis is a paraphrase of a subset of the text. For instance, given the text

Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pages 49–54,
Ann Arbor, June 2005.

Source	No. articles examined	No. articles in 1.5 mos.
miami-herald (US)	19	94,278
washington-post (US)	18	13,813
cs-monitor (US)	11	7,102
all-africa	18	68,521
dawn (Pakistan)	17	46,839
gulf-daily-news	10	26,837
national-post (Canada)	18	14,124

Figure 1: MiTAP News Sources Examined

John murdered Bill yesterday, the hypothesis *Bill is dead* is an entailment, while the hypothesis *Bill was killed by John* exhibits the stronger partial paraphrase relationship to the text. We found that 94% (131/140) of the TRUE pairs in the Pascal RTE dev2 corpus were these sorts of paraphrases.

In our search for an entailment corpus, we observed that the headline of a news article is often a partial paraphrase of the lead paragraph, much like the RTE data, or is sometimes a genuine entailment. We thus deduced that headlines and their corresponding lead paragraphs might provide a readily available source of training data. As an initial test of this hypothesis, we manually inspected over 200 news stories from 11 different sources. We found a great deal of variety in headline formats, and ultimately found the Xinhua News Agency English Service articles from the Gigaword corpus (Graff, 2003) to be the richest source, though somewhat limited in subject domain. We describe here our data collection and analysis process.

Because our goal was to automatically generate an extremely large corpus of exemplars, we focused on large data sources. We first examined 111 news stories culled from MiTAP (Damianos et al., 2003), which collects over one million articles per month from approximately 75 different sources. By first counting the number of articles typically collected for each source, we selected a mixture of sources that each had more than 10,000 articles for our sample period of one and half months. As discussed further below, part way through our investigation it became clear that we needed to include more native English sources, so the *Christian Science Monitor* articles were added, though they fell below our arbitrary 10K mark. Figure 1 summarizes the MiTAP news sources examined.

	Yes	**No**	**Maybe**	**Total**
All Pairs	54 (49%)	39 (35%)	18 (16%)	111
Filtered	54 (53%)	33 (33%)	14 (14%)	101

Figure 2: MiTAP Corpus Results

For each lead paragraph/headline pair, a human rendered a judgment of *yes*, *no,* or *maybe* as to whether the lead paragraph entailed the headline, where *maybe* meant that the headline was very close to being an entailment or paraphrase. This is likely equivalent to the notion of "more or less semantically equivalent" used in the Microsoft Research Paraphrase Corpus (Dolan et al., 2005). The purpose of *maybe* in this case was that we thought that many of the near-miss pairs would make adequate training data for statistical algorithms, in spite of being less than perfect.

There were many types of news articles in the MiTAP data that did not yield good headline/lead paragraph pairs for our purposes. Many would be difficult to filter out using automated heuristics. Two frequent examples of this were opinion-ed itorial pieces and daily Wall Street summaries. Others would be more amenable to automatic elimination, including obituaries and collections of news snippets like the *Washington Post*'s "World in Brief". Articles consisting of personal narratives never yielded good headlines, but these could easily be eliminated by recognizing first person pronouns in the lead paragraph. Figure 2 shows the judgments for all the MiTAP articles examined, where the Filtered row excludes these easily eliminated article types.

As Figure 2 shows, the MiTAP data did not yield a high percentage of good pairs. In addition, whether due to poor machine translation or English dialectal differences, our evaluator found it difficult to understand some of the text from sources that were not English-primary. A certain amount of ill-formed text was acceptable, since the Pascal RTE challenge included training and test data drawn from MT scenarios, but we did not wish our data to be too dominated by such sources. Thus, we selected additional native-English articles to add to our sample set.

Despite the overall poor yield from this data, it

Source	**Yes**	**No**	**Maybe**	**Total**
APW	8 (31%)	12 (46%)	6 (23%)	26
AFE	14 (56%)	4 (16%)	7 (28%)	25
NYT	8 (31%)	17 (65%)	1 (4%)	26
XIE	22 (85%)	4 (15%)	0 (0%)	26
Total	52 (50%)	37 (36%)	14 (14%)	103

Figure 3: Gigaword Corpus Results

was apparent that some news sources tended to be more fruitful than others. For example, 13 out of 18 of the *Washington Post* articles yielded good pairs, as opposed to only 1 of the 11 *Christian Science Monitor* articles.

This generalization was likewise true in the second corpus we examined, the Gigaword newswire corpus (Graff, 2003). Gigaword contains over 4 million documents from four news sources:

- Agence France Press English Service (AFE)
- Associated Press Worldstream English Service (APW)
- The New York Times Newswire Service (NYT)
- The Xinhua News Agency English Service (XIE)

For each source, Gigaword articles are classified into several types, including newswire advisories, etc. We restricted our investigations to actual news stories. As Figure 3 shows, overall results were much the same as the MiTAP articles, but 85% of the XIE articles yielded adequate pairs.

Based on these preliminary results we decided to focus further manual investigations on the XIE articles from Gigaword. We also decided to expend some effort on an annotation tool that would allow us to proceed more quickly than the early annotation experiments described above.

3 Refining the Data

MITRE has developed a series of annotation tools for a variety of linguistic phenomena (Day et al, 1997; Day et al, 2004), but these are primarily designed for fine-grained tasks such as named entity and syntactic annotation. For our headline corpus, we wanted the ability to rapidly annotate at a document level from a small set of categories. Further, we wanted the interface to easily support distributed annotation efforts.

The resulting annotation interface is shown in Figure 4. It is web-based, and annotations and other document information are stored in an SQL database. The document to be evaluated is displayed in the user's chosen browser, with the XML

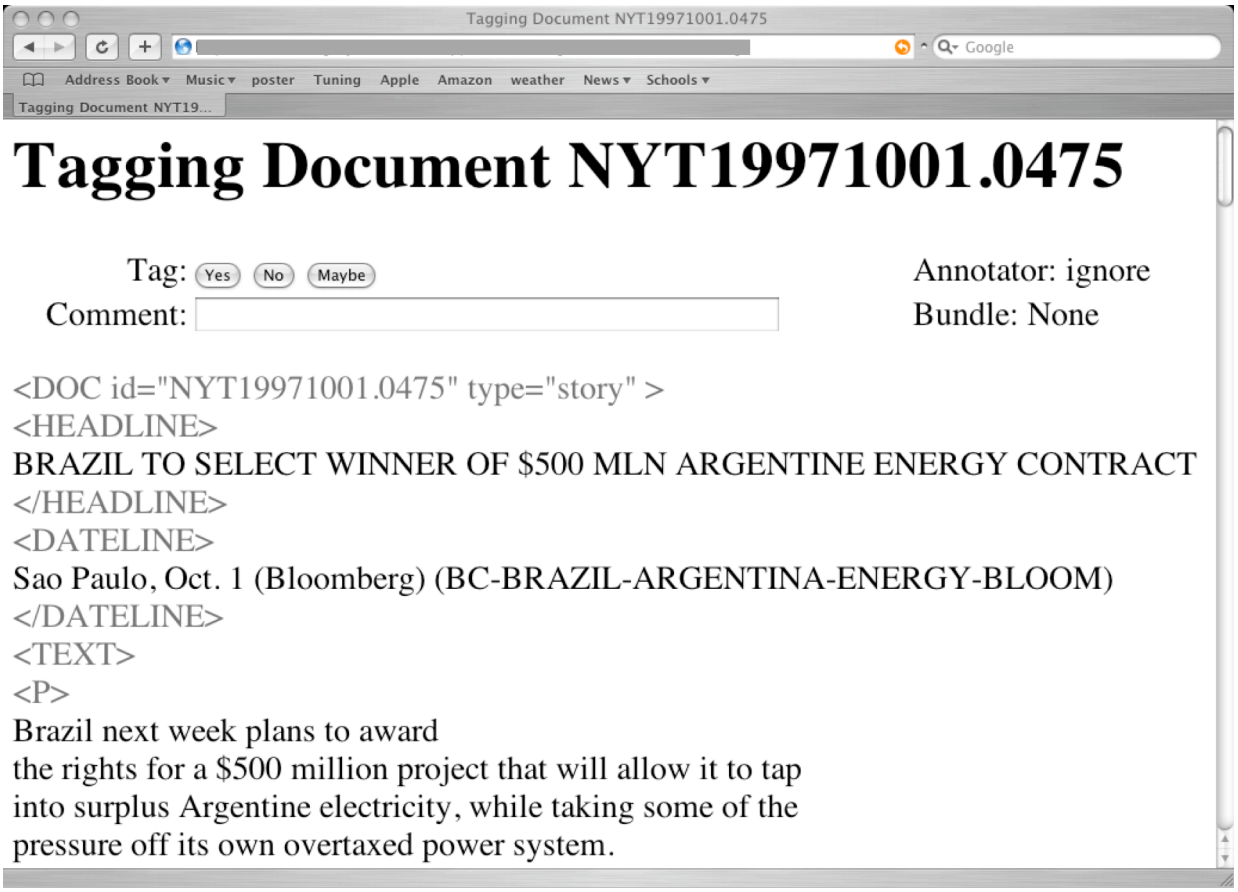

Figure 4: Entailment Tagging Interface

document zoning tags visible so that the user can easily identify the headline and lead paragraph. At the top of the document are three buttons from which to select a *yes/no/maybe* judgment. The user can also add a comment before moving to the next document. Typically several documents can be judged per minute. The client-server architecture supports multiple annotations of the same document by different annotators—accordingly, it has a mode enabling reconciliation of inter-annotator disagreements. All further annotation efforts discussed below were carried out with this tool.

Using the tool, we tagged approximately 900 randomly chosen Gigaword documents, including 520 XIE documents. From this, we estimate that 70% of the XIE headlines in Gigaword are entailed by the corresponding lead paragraph. (This is lower than the rough estimate described in Section 2, but that was based on a very small sample.) We decided to explore ways to refine the data in order to arrive at a smaller, but less noisy subcorpus. We observed that different subgenres within the newspaper corpus evinced the lead-entails-headline quality to different degrees. For example, articles about sports or entertainment often had whimsical (non-entailed) headlines, while articles about politics or business more frequently had the headline quality we sought.

Accordingly, we decided to treat the data refinement process as a text classification problem, one of finding the mix of genres or topics that would most likely possess the lead-entails-headline quality. We used SVM-light (Joachims, 2002) as a document classifier, training it on the initial set of annotated articles. (Note that these text classification experiments made use of the entire article, not just the lead and headline.) We experimented with a variety of feature representations and SVM parameters, but found the best performance with a Boolean bag-of-words representation, and a simple linear kernel. Leave-one-out estimates indicate that SVM-light could identify documents with the requisite entailment quality with 77% accuracy.

We performed one round of active learning (Tong & Koller, 2000), in which we used SVM-light to classify a large subset of the unannotated corpus, and then selected a 100-document subset about which the classifier was least certain. The rationale is that annotating these uncertain documents will be more informative to further learning runs than a randomly selected subset. In the case of large-margin classifiers like SVMs, the natural choice is to select the instances closest to the margin. These were then annotated, and added back to the training data for the next learning run. However, leave-one-out estimates indicated that the classifier benefited little from these new instances.

As described above, we estimate that the base rate of the headline entailment property in the XIE portion of Gigaword is 70%. Our hypothesis in training the SVM was that we could identify a smaller but less noisy subset. In order to evaluate this, we ran the trained SVM on all 679,000 of the unannotated XIE documents, and selected the 100,000 "best" instances—that is, the documents most likely (according to the SVM) to evince the headline quality. We selected a random subset of these best documents, and annotated them to evaluate our hypothesis. 74% of these possessed the lead-entails-headline property, a difference of 4% absolute over the XIE base rate. We used the lead-headline pairs from this 100,000-best subset to train our MT-alignment-based system for the RTE evaluation (Bayer et al., 2005). This system was one of the best performers in the evaluation, which we ascribe to our large training corpus

Later examination showed that the 4% "improvement" in purity is not statistically significant. We intend to perform further experiments in data refinement, but this may prove unnecessary. Perhaps the base rate of the entailment phenomenon in the XIE documents is sufficient to train an effective alignment-based entailment system. In this case, *all* of the XIE documents could be used, perhaps resulting in a more robust, and even better performing system.

4 Judging Headline Entailments

In the process of generating the training data, we doubly-judged an additional 300 XIE documents to measure inter-judge reliability. As in the pilot phase described above, each pair was labeled as *yes*, *no*, or *maybe*. In addition, the judges were given a comment field to record their reasoning and misgivings. The judging was performed in two steps, first on a set of 100 documents and then on a set of 200. One of the judges was already well versed in the RTE task, and had performed the earlier pilot investigations. Prior to judging the first set, the second judge was given a brief verbal

Condition	Set 1 (100 docs)	Set 2 (200 docs)
strict match	75.00%	77.50%
maybe = yes	79.00%	90.00%
maybe = no	84.00%	81.00%
maybe = *	88.00%	94.00%

Figure 5: Agreement for Two XIE Data Sets

overview of the task. After the first 100 documents had been doubly-judged, the more experienced judge then reviewed the differences and drafted a set of guidelines. The guidelines provided a synopsis of the official RTE guidelines, plus a few rules unique to headlines. For example, one rule specified what to do when partial entailment only held if the lead were combined with location or date information from the dateline. The two evaluators then judged the second set. The results for both sets are shown in Figure 5.

As these results show, the guidelines had only a small effect on the strict measure of agreement. Three problem areas existed:

(1) Raw, messy data. The Gigaword corpus was automatically collected and zoned. Thus, the headlines in particular contained a number of irregularities that made it difficult to judge their appropriateness. Such irregularities included truncations, phrases lacking any proposition, prepended alerts like *URGENT:,* and bylines and date lines miszoned into the headline.

(2) Disagreement on what constitutes synonymy. Our judges found they had irreconcilable differences about differences in meaning. For example, in the following pair, the judges disagreed about whether *safe operation* in the lead paragraph meant the same thing as, and thus entailed, *operates smoothly* in the headline:

- *Shanghai's Hongqiao Airport Operates Smoothly*
- *As of Saturday, Shanghai's Hongqiao Airport has performed safe operation for some 2,600 consecutive days, setting a record in the country.*

(3) Disagreement on the amount of world knowledge permitted. Figure 5 shows that if *maybe* is counted as equivalent to *yes*, the agreement level improves significantly. This is likely because there were two important aspects of the RTE definition of entailment that were not imparted to the second judge until the written guidelines: that one can assume "common human understanding of language and some common background knowledge." However, our judges did not always agree on what counts as "common," which accounts for much of the high overlap between *yes* and *maybe*. Nevertheless, our 90% agreement compares favorably to the 83% agreement rate reported by Dolan et al. (2005) for their judgments on "more or less semantically equivalent" pairs. Our 78% strict agreement compares favorably to the 80% agreement achieved by Dagan et al. (2005), given that our data was messier than the pairs crafted for the RTE challenge.

Like Dagan et al. (2005), we did not force resolution on all disagreements. Disagreements over synonymy and common knowledge result in irreconcilable differences, because it is neither possible nor desirable to use guidelines to force a shared understanding of an utterance. Thus, for the first set of data 15 (15%) of the pairs were left unreconciled. In the second set, 42 (21%) were left unreconciled. Eleven (6%) of the irreconcilable pairs in the second set were due to confusion stemming from the telegraphic nature of headlines, which led to misunderstandings about how to judge truncated headlines (*Chinese President Vows to Open New Chapters With*) vs. headlines lacking propositions (subject headings like *Mandela's Speech*) vs. well-formed but terse headlines (*Crackdown on Auto-Mafia in Bulgaria*).

Despite the high number of irreconcilable pairs, one encouraging sign was evident from the comment field. The judges' comments revealed that on pairs where they disagreed on how to label the pair, they often agreed on what the problem was.

Our experience in generating a training corpus, particularly the number of irreconcilable cases we encountered, raises an important issue, namely, the feasibility of semantic equivalence tasks. We suggest that the optimum method for empirically modeling semantic equivalence is to capture the variation in human judgments. Three judges would evaluate each pair, so that there would always be a tie breaker. After reconciling for disagreements arising from human error, each distinct judgment would become part of the data set. We also recommend that where there is genuine disagreement, the questionable portions of each pair be annotated in some way to capture the source of the problem, going one step further than the comment field we found beneficial in our annotation interface. The three judgments would result in a four way classification of pairs:

TTT = TRUE
TTF = Likely TRUE, but possibly FALSE
TFF = Likely FALSE, but possibly TRUE
FFF = FALSE

System developers could choose to train on all the data, or limit themselves to the TTT/FFF cases. For evaluation purposes, the systems' results on the TTF/TFF pairs could be evaluated in light of the human variation, providing a more realistic measure of the complexity of the task.

5 Conclusion

Given the number of natural language processing applications that require the ability to recognize semantic equivalence and entailment, there is an obvious need for both robust evaluation methodologies and adequate development and test data. We've described here our work in generating supplemental training data for the recent Pascal RTE evaluation, with which we produced a competitive system. Some news corpora provide a rich source of exemplars, and an automatic document classifier can be used to reduce the noisiness of the data. There are lingering difficulties in achieving high inter-judge agreement in determining paraphrase and entailment, and we believe the best way to cope with this is to allow the data to reflect the variance that exists in cross-human judgments.

Acknowledgments

This paper reports on work supported by the MITRE Sponsored Research Program. We would also like to extend our thanks to Sam Bayer, John Henderson and Alex Yeh for their invaluable suggestions and comments. Our gratitude also goes to Laurie Damianos, who provided us with statistics on MiTAP's resources and served as one of the evaluators in our inter-judge reliability study.

References

Samuel Bayer, John Burger, Lisa Ferro, John Henderson, and Alexander Yeh, 2005. MITRE's submissions to the EU Pascal RTE Challenge. *PASCAL Proceedings of the First Challenge Workshop, Recognizing Textual Entailment*, 11–13 April, 2005, Southampton, U.K.

John D. Burger, 2004. MITRE's Qanda at TREC-12. *The Twelvth Text REtrieval Conference*. NIST Special Publication SP 500–255.

Ido Dagan, Oren Glickman, and Bernado Magnini, 2005. The PASCAL recognizing textual entailment challenge. *PASCAL Proceedings of the First Challenge Workshop, Recognizing Textual Entailment*, 11–13 April, 2005, Southampton, U.K.

Laurie Damianos, Steve Wohlever, Robyn Kozierok, and Jay Ponte, 2003. mitap for real users, real data, real problems. In *Proceedings of the Conference on Human Factors of Computing Systems* (CHI 2003), Fort Lauderdale, FL April 5–10.

David Day, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson and Marc Vilain, 1997. Mixed initiative development of language processing systems. *Proceedings of the Fifth Conference on Applied Natural Language Processing*.

David Day, Chad McHenry, Robyn Kozierok, Laurel Riek, 2004. Callisto: A configurable annotation workbench. *International Conference on Language Resources and Evaluation*.

Bill Dolan, Chris Brockett., and Chris Quirk, 2005. *Microsoft Research Paraphrase Corpus*. http://research.microsoft.com/research/nlp/msr_paraphrase.htm

David Graff, 2003. *English Gigaword*. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05

Dan Gusfield, 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press.

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger, 1999. Deep Read: A reading comprehension system. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*.

Thorsten Joachims, 2002. *Learning to Classify Text Using Support Vector Machines*. Kluwer.

Guido Minnen, John Carroll, and Darren Pearce, 2001. Applied morphological processing of English. *Natural Language Engineering,* 7(3).

Franz Josef Och and Hermann Ney, 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1).

Simon Tong and Daphne Koller, 2000. support vector machine active learning with applications to text classification. *Proceedings of ICML-00, 17th International Conference on Machine Learning*.

Ben Wellner, Lisa Ferro, Warren Greiff, and Lynette Hirschman, 2005. Reading comprehension tests for computer-based understanding evaluation. *Natural Language Engineering* (to appear).